arXiv:2301.11477v1 [stat.ME] 27 Jan 2023

Ananke：使用图模型进行因果推断的 Python 软件包

JaronJ. R.Lee[†, 1] JARON.LEE@JHU.EDU
Rohit Bhattacharya[†, 2] RB17@WILLIAMS.EDU
Razieh Nabi[3] RAZIEH.NABI@EMORY.EDU
Ilya Shpitser[1] ILYAS@CS.JHU.EDU
† **共同贡献**[1]
约翰斯·霍普金斯大学计算机科学系[2]威廉姆斯学院计算机科学系
[3]埃默里大学生物统计学与生物信息学系

摘要

我们实现了Ananke：一个用于基于图模型进行因果推断的面向对象 Python 软件包。继承结构的顶层是一个易于扩展的图类，为多种广泛适用的基于图的算法及可视化方法提供接口。我们遵循面向对象编程的最佳实践，实现了图超类的子类，分别对应当前文献中常见的因果图类型。这些类型包括用于对因果充分系统建模的有向无环图、用于对未测量混杂建模的无环有向混合图，以及用于对数据依赖和干扰建模的链图。在这些子类中，我们实现了常见统计与因果建模任务所需的专用算法，例如用于读取条件独立关系的分离准则、非参数识别，以及模型参数的参数估计和半参数估计。本文概述了该软件包，并通过一个存在未测量混杂的问题展示其用法。最新文档请见 https://ananke.readthedocs.io/en/latest/。

关键词： 因果图模型、因果识别、半参数估计

1. 引言

因果推断是一套由多个步骤组成的流程——指定因果模型、在该模型的假设下识别所需的因果参数、根据识别函数利用数据估计该参数，以及通过敏感性分析和不确定性量化进行稳健性检验。未测量混杂、数据依赖和缺失数据都可能使其中任何一步变得复杂。在 Ananke 中，我们实现了涵盖所有这些步骤的方法，包括非参数识别和半参数估计策略，为分析人员提供统一的接口，使其能够建立端到端的流程，并增强流程对上述复杂情况的稳健性。

具体而言，我们采用面向对象范式来实现基于图的因果推断方法。我们构建了一套继承结构，涵盖采用任意组合的有向（$\rightarrow$）、双向（$\leftrightarrow$）和无向（$-$）边的因果图模型。我们希望，凭借其面向对象的特性和易于使用的 Python 实现

©2022 Lee、Bhattacharya、Nabi 和 Shpitser。许可协议：知识共享署名 4.0 国际许可协议，参见 https://creativecommons.org/licenses/by/4.0/。

Ananke 将提高许多基于图的因果推断方法的易用性，并使感兴趣的用户能够轻松扩展和构建其现有基础设施。

相关工作：doWhy 包和 DAGitty 旨在为因果推断流程中的各个不同步骤提供统一的接口。然而，它们的估计能力在很大程度上仅限于不存在未测量混杂因素或选择偏倚的情形。对于 doWhy，其因果图接口仍在积极开发中，计划与 Ananke 对接，而不是从头构建（与开发者的个人交流）。其他现有软件包则侧重于流程中的单个步骤。TETRAD 及其 Python 移植版 causal-learn（Scheines 等，1998）、pcalg（Kalisch 等，2012）和 cdt（Kalainathan 等，2020）侧重于图表示和模型选择；causaleffect 侧重于非参数识别；npcausal（Kennedy，2021）、zEpid（Zevich，2018）、tmle3（Coyle，2021）和 DoubleML（Bach 等，2022）侧重于半参数估计。其他独立软件包作为论文附录发布；在某些情况下，我们会在 Ananke 中重新实现这些算法，例如 Evans（2013）和 Drton 等（2009）中的最大似然算法。与同类软件包相比，Ananke 的主要优势在于，它在单个软件包中为因果推断提供了统一且易于扩展的接口，并拥有一个活跃的社区。

2. Ananke的图继承结构概览

图模型概览Ananke及其继承结构如图 1 所示。当前，图 类支持创建图 $\mathcal{G} = (V, D, B, U)$，其中 V 表示顶点集合，D、B、U 分别表示有向（$\rightarrow$）、双向（$\leftrightarrow$）和无向（$-$）边的集合。在此类中，我们实现了广泛适用于任何子类的方法和算法：包括添加和删除边、查找仅由 $\mathcal{G}_S$ 中的顶点组成的子图 $S \subseteq V$（以及关联

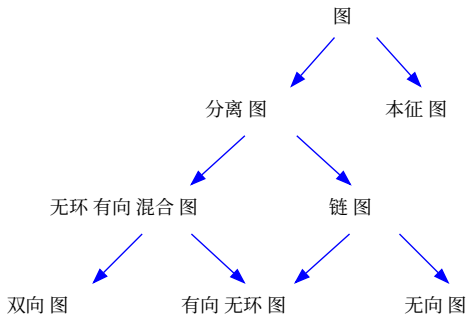


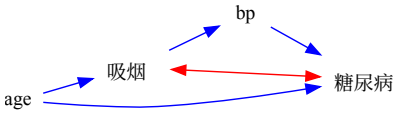
图 1：Ananke 的图继承结构。

边），以及计算顶点的谱系集合，例如其祖先、后代和兄弟节点。我们还通过 Python 接口实现了一个轻量级的绘图方法，使用Graphviz（Ellson 等人，2001；Hagberg 等人，2022）来可视化该类或其子类的任意实例——本文中的所有图均使用此功能生成。其余继承结构则依据各图类所包含的边类型构建。层级最低的是仅包含一种边类型的图：有向无环图（DAG）（仅包含 $\rightarrow$边）是最常见的因果图类型（Robins，1986；Spirtes 等人，2000；Pearl，2009）；双向图（仅包含 $\leftrightarrow$ 边）用于表示边际相关，在基因组学中十分常见（Chaudhuri 等人，2007；Cox 和 Wermuth，2014）；无向图（仅包含$-$ 边）可用于编码反馈关系（Lauritzen，1996）。接下来是包含多种边类型的图：无环有向混合图（ADMG）用于对具有因果影响的系统（通过 $\rightarrow$ 边）以及由未测量混杂导致相关性的系统（通过 $\leftrightarrow$ 边）进行建模（Wright，1921；Verma 和 Pearl，1990）；链图则用于对因果影响（通过 $\rightarrow$ 边）以及诸如传染等非独立同分布现象进行建模——

、反馈和对称关系（通过– 边）（Lauritzen 和 Richardson，2002；Ogburn 等人，2020；Bhattacharya 等人，2019a）；由三种边构成的分离图能够对上述三种机制进行建模（Shpitser，2015）。需要指出的是，图 1 层级结构中展示的内在图并非因果图形模型，而是我们创建的一种图形表示，用于高效计算参数化隐变量因果模型所需的全部统计核——这是第 3 节所讨论估计过程中的必要步骤。这展示了我们的图继承结构在中间任务中的其他应用场景。举例来说，我们使用链图实现来编码因果有向无环图的等价类——即对观测数据分布施加相同限制的不同模型——这类图称为完全部分有向无环图（CPDAG）。这样，`Ananke` 便可轻松接入或扩展能够输出此类对象的因果发现算法，例如 `causal-learn` 包中的贪婪等价搜索或 PC 算法实现（Zhang 等人，2022）。

3. 数据分析（ `Ananke`）

为了说明 Ananke 的用法，我们将逐步介绍一项假设性分析，评估吸烟对糖尿病的效应。该分析使用一个教学数据集，该数据集源自弗雷明汉心脏研究（Kannel 和 Gordon，1968）[1]。我们首先使用图 2 所示的祖先有向混合图（ADMG）来编码实质性假设，并给出用于创建和可视化该图的 Ananke 命令。




```
>>>顶点 = ["年龄",      "吸烟",   "血压", "糖尿病"]
>>>二_条边 = [("年龄",   "吸烟"),   ("吸烟",   "血压"),
>>>                ("bp", "糖尿病"),    ("年龄", "糖尿病")]
>>>双_边 = [("吸烟",      "糖尿病")]
>>> fdoor = 图.祖先有向混合图(vertices, di_edges, bi_edges)
>>> fdoor.draw("LR")
```

图 2：使用内置功能可视化的前门模型。

祖先有向混合图可以蕴含某些可检验的独立性陈述（以及其他一般约束），这些陈述可通过 m-分离读出（Richardson，2003）。例如，我们可以验证，在给定吸烟的条件下，年龄与血压是 m-分离的，这意味着年龄 ⫫ 血压 | 吸烟。

```
>>> fdoor.m_分离("年龄","血压", ["吸烟"])真
```

分析人员可以使用任何标准的条件独立性检验，验证数据是否支持此类假设。假设图 2 正确，下一步是应用识别理论，以确定所需的因果效应能否表示为观测数据的函数。使用 Ananke 实现的、在存在未测量混杂时可靠且完备的识别算法（Richardson 等人，2017），结果如下：

```
>>> 处理 = ["吸烟"]; 结果 = ["糖尿病"]>>> 识别_fdoor =识别。OneLineID(图=fdoor, 处理=treatments, 结果=outcomes)>>> print('已识别 =', id_fdoor.id(), '; 函数形式 =', id_fdoor.functional()) 已识别 = True 函数形式 = Σ年龄血压 𝜙糖尿病吸烟血压(p(V);G) 𝜙吸烟糖尿病年龄(p(V);G) 𝜙吸烟血压年龄(p(V);G)
```

1. 弗雷明汉心脏研究的教学摘录可通过 https://biolincc.nhlbi. nih.gov/teaching/ 申请。

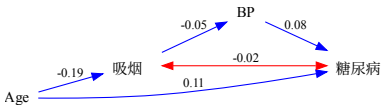
也就是说，反事实分布 $p(\text{糖尿病(吸烟)})$，以及由此得到的效应，确实是可识别的。根据 Richardson 等（2017）对输出进行解读，可得

$$p(\texttt{diabetes}(\texttt{smoke})) = \sum_{\texttt{age},\texttt{bp}} p(\texttt{bp} \mid \texttt{smoke}) p(\texttt{age}) \left(\sum_{\texttt{smoke}'} p(\texttt{diabetes} \mid \texttt{smoke}', \texttt{bp}, \texttt{age}) p(\texttt{smoke}' \mid \texttt{age}) \right).$$

虽然本分析侧重于未测量混杂条件下的识别，但 Ananke 也实现了针对缺失数据（Bhattacharya 等人，2019b；Nabi 等人，2020）、选择偏倚和数据融合（Lee 等人，2020；Lee 和 Shpitser，2020）的识别算法。完成识别后，我们可以从 Ananke 提供的多种估计策略中进行选择。

3.1 线性高斯结构方程模型

一种可能的选择是假设误差相关的线性结构方程模型（Wright，1934）。在 Ananke 中，我们实现了 Drton 等（2009）所述的迭代算法，以获得所有边系数的最大似然估计；随后通过路径分析计算因果效应。将此方法应用于标准化的 Framingham 数据，得到：




```
>>> lsm  = LinearGaussianSEM(front_door)>>> lsm  = lsm.fit(df_cont)>>> lsm.draw(direction="LR")>>> lsm.total_effect(["smoke"], ["diabetes"])平均因果效应：-0.004
```

3.2 离散数据的 Möbius 参数化

另一种方法是使用观测数据似然的 Möbius 参数化，该方法假设所有观测变量均为离散变量（Evans 和 Richardson，2012）。我们实现了一种坐标下降算法，用于计算 Möbius 参数的最大似然估计（一般而言，这些参数可能存在变分依赖关系）。使用数据集中的变量的二元版本，我们得到平均因果效应如下。

```
>>> bnm  =binary_nested.BinaryNestedModel(front_door)>>> bnm  =bnm.fit(X=binary_nested.process_data(df_bin), tol=1e-12 )>>> pY1_A0  =bnm.estimate(treatment_dict={"smoke": 0}, outcome_dict={"diabetes": 1})>>> pY1_A1  =bnm.estimate(treatment_dict={"smoke": 1}, outcome_dict={"diabetes": 1})>>> print("ACE: ", pY1_A1 - pY1_A0) 平均因果效应：0.004
```

3.3 因果效应的半参数估计

如果效应可识别，Ananke 会列出 Bhattacharya 等（2022）提出的若干半参数估计策略，并根据半参数效率理论推荐最佳策略。该实现只需指定祖先有向混合图、处理和结果。采用 Ananke 推荐的高效增广原始 *IPW* 估计量可得：

```
>>> 平均因果效应_对象  =因果效应（图=前门_，处理='吸烟'，结果='糖尿病'）>>> 平均因果效应  =平均因果效应_对象。计算_效应(df_bin,"eff-apipw"); 打印("平均因果效应：", 平均因果效应)平均因果效应：-0.002
```

致谢

感谢 PreethiPrakash 和 Ranjani Srinivasan 对 Ananke 的贡献，并感谢 Carson Kurtz 协助 R.B. 测试代码。

参考文献

Philipp Bach、Victor Chernozhukov、Malte S. Kurz 和 Martin Spindler。DoubleML——一种用 Python 实现双重机器学习的面向对象方案。机器学习研究杂志，23(53):1–6，2022。网址：http://jmlr.org/papers/v23/ 21-0862.html。

Rohit Bhattacharya、Daniel Malinsky 和 Ilya Shpitser。干扰与网络不确定性下的因果推断。载于第*35*届人工智能不确定性会议论文集。AUAI出版社，2019a。

Rohit Bhattacharya、Razieh Nabi、Ilya Shpitser 和 James M. Robins。有向无环图表示的缺失数据模型中的识别。载于第*35*届人工智能不确定性会议论文集。AUAI出版社，2019b。

Rohit Bhattacharya、Razieh Nabi 和 Ilya Shpitser。含隐变量图模型中因果效应的半参数推断。机器学习研究杂志，23:1–76，2022。

Sanjay Chaudhuri、Mathias Drton 和 Thomas S Richardson。含零元素协方差矩阵的估计。*Biometrika*，94(1)：199–216，2007。

David Roxbee Cox 和 Nanny Wermuth。多元依赖关系：模型、分析与解释。Chapman and Hall/CRC，2014。

Jeremy R Coyle。tmle3：可扩展的 TMLE 框架。https://github.com/tlverse/ tmle3，2021。URL https://doi.org/10.5281/zenodo.4603358。R 软件包版本 0.2.0。

Mathias Drton、Michael Eichler 和 Thomas S. Richardson。在具有相关误差的递归线性模型中计算最大似然估计。机器学习研究杂志，10(10)，2009。

John Ellson、Emden Gansner、Lefteris Koutsofios、Stephen C North 和 Gordon Wood-hull。Graphviz——开源图绘制工具。载于国际图绘制研讨会，第 483–484 页。施普林格，2001。

Robin J Evans。ADMG。https://www.stats.ox.ac.uk/~evans/software.htm，2013。

Robin J. Evans 和 Thomas S. Richardson。将最大似然拟合应用于二元数据的无环有向混合图。2012。网址：http://arxiv.org/abs/1203.3479。

Aric Hagberg、Dan Schult 和 Manos Renieris。Pygraphviz。2022。网址：https://github. com/pygraphviz/pygraphviz。

Diviyan Kalainathan、Olivier Goudet 和 Ritik Dutta。因果发现工具箱：使用 Python 揭示因果关系。机器学习研究杂志，21(37):1–5，2020。Markus Kalisch、Martin M¨achler、Diego Colombo、Marloes H. Maathuis 和 Peter B¨uhlmann。使用 R 软件包 pcalg 通过图模型进行因果推断。统计软件杂志，47:1–26，2012。William B. Kannel 和 Tavia Gordon。弗雷明汉研究：一项关于心血管疾病的流行病学调查。第 9-13 号。卫生、教育与福利部，1968。Ed Kennedy。npcausal：非参数因果推断方法。https://github.com/ ehkennedy/npcausal，2021。Steffen L. Lauritzen。图模型。英国牛津：克拉伦登，1996。Steffen L. Lauritzen 和 Thomas S. Richardson。链图模型及其因果解释。英国皇家统计学会杂志：*B*辑（统计方法学），64(3):321–348，2002。Jaron J. R. Lee 和 Ilya Shpitser。以任意干预分布为输入的识别方法。*arXiv* 预印本 *arXiv:2004.01157*，2020。Sanghack Lee、Juan D Correa 和 Elias Bareinboim。通过任意替代实验实现一般可识别性。载于人工智能中的不确定性，第 389–398 页。PMLR，2020。Razieh Nabi、Rohit Bhattacharya 和 Ilya Shpitser。缺失数据图模型中的完整分布识别：完备性结果。载于国际机器学习会议，第 7153–7163 页。PMLR，2020。Elizabeth L. Ogburn、Ilya Shpitser 和 Youjin Lee。因果推断、社会网络与链图。英国皇家统计学会杂志：*A*辑（社会统计学），183(4):1659–1676，2020。Judea Pearl。因果关系。剑桥大学出版社，2009。Thomas S. Richardson。无环有向混合图的马尔可夫性质。斯堪的纳维亚统计学杂志，30(1):145–157，2003。Thomas S. Richardson、Robin J. Evans、James M. Robins 和 Ilya Shpitser。无环有向混合图的嵌套马尔可夫性质，2017。工作论文。James M. Robins。一种研究持续暴露期间死亡率因果推断的新方法——应用于控制健康工人幸存者效应。数学建模，7(9-12):1393–1512，1986。Richard Scheines、Peter Spirtes、Clark Glymour、Christopher Meek 和 Thomas Richardson。Tetrad 项目：用于因果模型设定的基于约束的辅助工具。多变量行为研究，33(1):65–117，1998。

Ilya Shpitser。《链图模型的分离图与边缘分布》。载于神经信息处理系统进展，第1720–1728页，2015年。Peter L. Spirtes、Clark N. Glymour 和 Richard Scheines。《因果、预测与搜索》。麻省理工学院出版社，2000年。Thomas Verma 和 Judea Pearl。因果模型的等价性与合成。载于第六届人工智能不确定性年度会议论文集，1990年。Sewall Wright。相关性与因果关系。《农业研究杂志》，20：557–580，1921年。Sewall Wright。路径系数法。《数理统计年刊》，5(3)：161–215，1934年。Paul Zevich。zEpid：流行病学分析工具包。https://github.com/pzivich/ zepid/，2018年。Python 软件包。Kun Zhang、Joseph Ramsey、Mingming Gong、Ruichu Cai、Shohei Shimizu、Peter Spirtes、Clark Glymour 等。causal-learn：用于 Python 的因果发现。https://github.com/cmu-phil/causal-learn，2022年。